Sequential Analysis with More than Two Alternative Hypotheses, and its Relation to Discriminant Function Analysis

By P. Armitage

Medical Research Council Statistical Research Unit, London School of Hygiene and Tropical Medicine

[Received November 11th, 1949]

1. *Introduction*

The theory of sequential tests for deciding between two alternative simple hypotheses is now well known, and is described by Wald (1947). In a review of Wald's book, Barnard (1947) pointed out that, as a generalization of Wald's method, one could formulate a procedure for deciding between more than two simple hypotheses. The procedure would be based on likelihood ratios, and the risks involved could be controlled by suitable choice of the acceptance conditions.

The present note contains an outline of the theory of such sequential procedures, which are closely related to some recent developments in the theory of discriminant functions. The methods should prove useful for two-sided tests of statistical hypotheses. In particular, I consider a two-sided test for the value of a binomial probability, and, as a development of this, a two-sided test for comparative trials.

Since the present paper was submitted for publication Sobel and Wald (1949) have published details of a sequential decision procedure for choosing one of three mutually exclusive and exhaustive composite hypotheses about the mean of a normal distribution. Their procedure involves the combination of two tests for distinguishing between two pairs of simple hypotheses, and is closely related to that suggested by Armitage (1947) for a two-sided sequential t-test. The discussion at the end of §4 of the present paper has been added since the publication of Sobel and Wald's paper.

2. *Theory*

Consider k simple hypotheses, $H_1, H_2, \ldots, H_k$, and let the likelihood from a single observation, when H_i is true, be L_i. There are $\frac{1}{2}k(k-1)$ likelihood ratios for various pairs of hypotheses, but each of these may be expressed in terms of $k-1$ independent likelihood ratios, which may be chosen in any one of a number of different ways. We could take, for instance,

$$R_i = L_i/L_k, \quad i = 1, 2, \ldots, k - 1,$$

and let

$$y_i = \log R_i.$$

(The base of the logarithms is arbitrary.) Then the logarithm of the likelihood ratio for any two hypotheses is either one of the y_i, or a difference between two of the y_i.

Successive observations are taken to be independent, and the logarithm of the likelihood ratio for two of the hypotheses, after n observations, will be of the form

$$\Sigma\, y_i \text{ or } \Sigma\, (y_i - y_j),$$

the summation being over the n observations.

Let us formulate the following rule of procedure. The observations are taken sequentially, until all the inequalities in one of the following k sets are simultaneously satisfied:

$$\left.\begin{array}{ll}\Sigma(y_i - y_1) & > A_{i,1} \\ \quad\vdots & \\ \Sigma(y_i - y_{i-1}) & > A_{i,i-1} \\ \Sigma(y_i - y_{i+1}) & > A_{i,i+1} \\ \quad\vdots & \\ \Sigma(y_i - y_{k-1}) & > A_{i,k-1} \\ \Sigma y_i & > A_{i,k}\end{array}\right\}\ \text{in which case } H_i \text{ is accepted, } (i = 1, 2, \ldots, k-1) \quad . \quad . \quad . \quad (1)$$

or

$$\left.\begin{array}{ll}\Sigma y_1 & < -A_{k,1} \\ \quad\vdots & \\ \Sigma y_{k-1} & < -A_{k,k-1}\end{array}\right\}\ \text{in which case } H_k \text{ is accepted} \quad . \quad . \quad . \quad (2)$$

the $A_{i,j}$ being arbitrary positive constants.

The constants $A_{i,j}$ could conveniently be made equal, say $A_{i,j} = A$. In this case the inequalities specify that sampling continues until the likelihood ratios of one of the hypotheses against *each* of the others are all greater than antilog A. (This procedure was suggested by Barnard (1947)). For generality we retain different values of $A_{i,j}$.

We shall now show that, if a sequential procedure is defined by the inequalities (1) and (2), the probability that no decision has been reached by the n^{th} stage tends to zero as n increases indefinitely. This result is true under hypotheses which are not necessarily confined to the set $H_1, \ldots, H_k$, provided that the distributions of the likelihood functions y_i satisfy certain regularity conditions.

If, at the n^{th} stage, no decision has been reached, then, at each stage up to and including the n^{th}, at least one of the following ${}^kC_2$ relations must have been satisfied:

$$\begin{array}{ll} -A_{k,i} \leqslant \Sigma y_i \leqslant A_{i,k} & (i = 1, \ldots, k-1) \\ -A_{j,i} \leqslant \Sigma(y_i - y_j) \leqslant A_{i,j} & (i, j = 1, \ldots, k-1) \quad . \quad . \quad . \quad (3)\end{array}$$

The probability that this condition is satisfied is less than the probability that at least one of the inequalities (3) is satisfied at the n^{th} stage, irrespective of the previous stages. Provided the variances of the distributions of y_i and $(y_i - y_j)$ all exist, we can apply the central limit theorem to the distribution of the sums Σy_i and $\Sigma(y_i - y_j)$. It follows that the probability that a given inequality of the set (3) is satisfied at the n^{th} stage can be made as small as we wish by a suitable choice of n. The probability that at least one of the inequalities (3) is satisfied at the n^{th} stage therefore tends to zero as n tends to infinity, and finally the probability that the procedure has terminated by the n^{th} stage tends to unity as n tends to infinity.

Let π_{ij} be the probability of accepting H_i when in fact H_j is true. By considering the total probability of all samples which result in H_i being accepted, we see from (1) and (2) that

$$1 > \pi_{ii} > B_{i,j}\,\pi_{ij} \qquad (i, j = 1, 2, \ldots k, i \neq j),$$

where $B_{i,j} =$ antilog $A_{i,j}$. Hence

$$\pi_{ij} < 1/B_{i,j} \qquad (i \neq j).$$

It follows that

$$\pi_{ii} = 1 - \sum_{j \neq i} \pi_{ji}$$

$$> 1 - \sum_{j \neq i} 1/B_{j,i} \qquad (i = 1, 2, \ldots, k) \quad . \quad . \quad . \quad (4)$$

The inequalities (4) may be used to control the risks of error associated with any sequential procedure. By choosing the $A_{i,j}$ sufficiently large we can make the probabilities of arriving at the correct conclusion, when any one of the H_i is true, as large as we wish. These inequalities are, however, conservative, in the sense that the true probabilities may be considerably higher than the lower bound given by (4). As we shall see below, it is possible in certain problems to assert that certain of the π_{ij} are effectively zero, and so improve on the inequalities (4).

3. *Application to Discriminant Function Analysis*

Welch (1939) and, more recently, Smith (1947) have pointed out that Fisher's linear function for discriminating between two multivariate normal populations, with different means but the same covariance matrix, is equivalent to the likelihood ratio between the two populations. Rao (1948) has used this fact in developing a method of discrimination between $k > 2$ populations. Apart from an adjustment for *a priori* knowledge, Rao's method consists in assigning an individual to the i^{th} group if (in our notation)

$$\left.\begin{array}{ll} (y_i - y_1) & > 0 \\ \cdot & \\ \cdot & \\ \cdot & \\ (y_i - y_{i-1}) & > 0 \\ (y_i - y_{i+1}) & > 0 \\ \cdot & \\ \cdot & \\ \cdot & \\ y_i & > 0 \end{array}\right\} (i = 1, 2, \ldots, k-1), \qquad \left.\begin{array}{ll} y_1 & < 0 \\ \cdot & \\ \cdot & \\ \cdot & \\ y_{k-1} & < 0 \end{array}\right\} (i = k).$$

In other words, the individual is assigned to the most likely population. The probabilities of correctly or incorrectly classifying an individual from any one group can be calculated by means of the probability integral of the multivariate normal distribution of the y_i, which is tabulated for $k = 2$ and $k = 3$.*

Suppose we have a group of individuals known to come from one or other of the k populations, and we wish to decide which one this is. The problem is clearly equivalent to that discussed in §2, and the rules of procedure for a sequential method are given by (1) and (2). The risks of incorrect classification are controlled by (4). This sequential procedure has in fact already been suggested by Rao (1947), for the case $k = 2$.

As an illustration we may consider the case $k = 3$, with discriminant functions

$$y_1 = \log (L_1/L_3),$$

$$y_2 = \log (L_2/L_3).$$

In the standard method described by Rao we should assign an individual to populations 1, 2 or 3 according to which region of Fig. 1 the point (y_1, y_2) lay in.

In the sequential method we continue to sample until the point $(\Sigma\, y_1, \Sigma\, y_2)$ first falls in one of the three shaded areas of Fig. 2.

If, in particular, we choose $A_{ij} = A$ for $i, j = 1, 2, 3$, we find from (4) that the probabilities that an individual from each of the three populations will be correctly classified are all greater

* For $k = 3$ there are two discriminant functions, and the integral of the bivariate normal surface may be obtained from Tables VIII and IX of Pearson (1931).

FIG. 1.

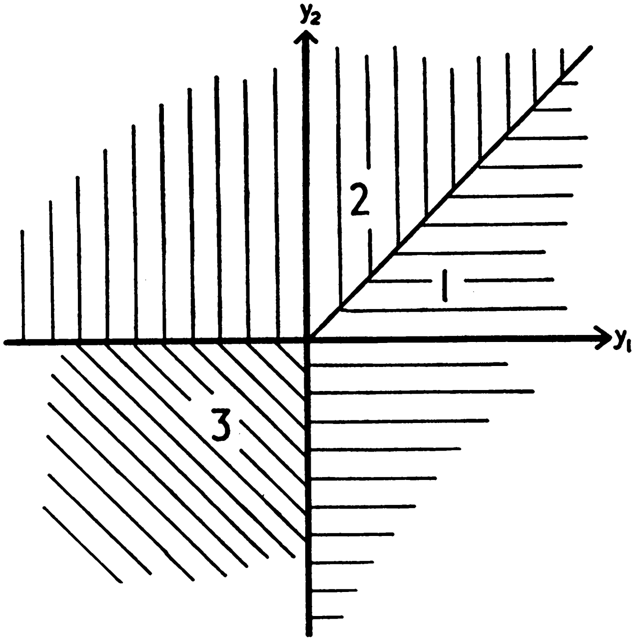
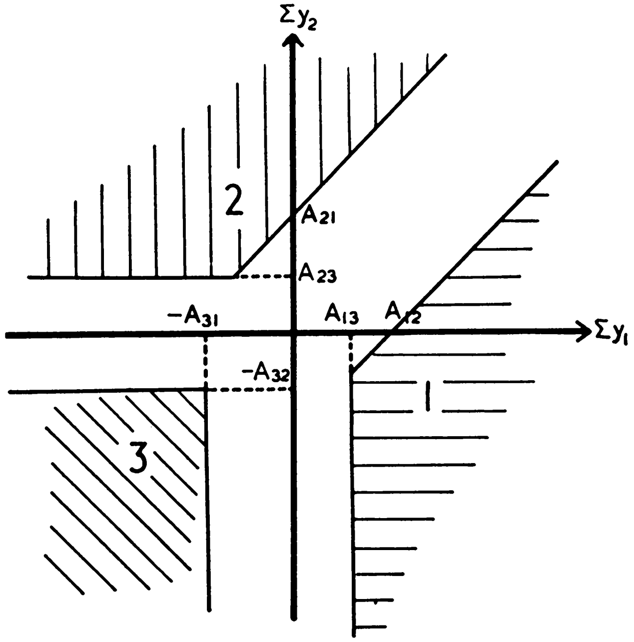


FIG. 2.



FIGS. 1 and 2.—Discrimination by non-sequential and sequential methods.

than $1 - 2/(\text{antilog } A)$. By choosing A sufficiently large, and the boundaries of the three regions sufficiently far from the origin, these probabilities may be brought as near to unity as we please.

4. *Test for Three Binomial Probabilities*

Denote by p the probability of occurrence of a certain event in each of a series of independent trials, and let H_1, H_2 and H_3 be the hypotheses that p takes the values p_1, p_2 and p_3 respectively, where $p_1 < p_2 < p_3$. Suppose we wish to decide between H_1, H_2 and H_3. If, after n trials, the event occurred m times, we have—

$$\left.\begin{aligned} \Sigma y_1 &= \Sigma \log (L_1/L_3) = m \log (p_1/p_3) - (n - m) \log (1 - p_1)/(1 - p_3) \\ \text{and} \quad \Sigma y_2 &= \Sigma \log (L_2/L_3) = m \log (p_2/p_3) - (n - m) \log (1 - p_2)/(1 - p_3) \end{aligned}\right\} \quad . \quad . \quad (5)$$

the summations, as usual, being over the n trials.

By (1) and (2) we accept H_1 at the n^{th} stage if $\Sigma (y_1 - y_2) > A_{12}$ and $\Sigma y_1 > A_{13}$; we accept H_2 if $\Sigma (y_1 - y_2) < - A_{21}$ and $\Sigma y_2 > A_{23}$; and we accept H_3 if $\Sigma y_1 < - A_{31}$ and $\Sigma y_2 < - A_{32}$.

These conditions may be seen to lead to a graphical procedure of plotting the point $(n - m, m)$ on Barnard's inspection diagram,* with boundaries as shown in Fig. 3.

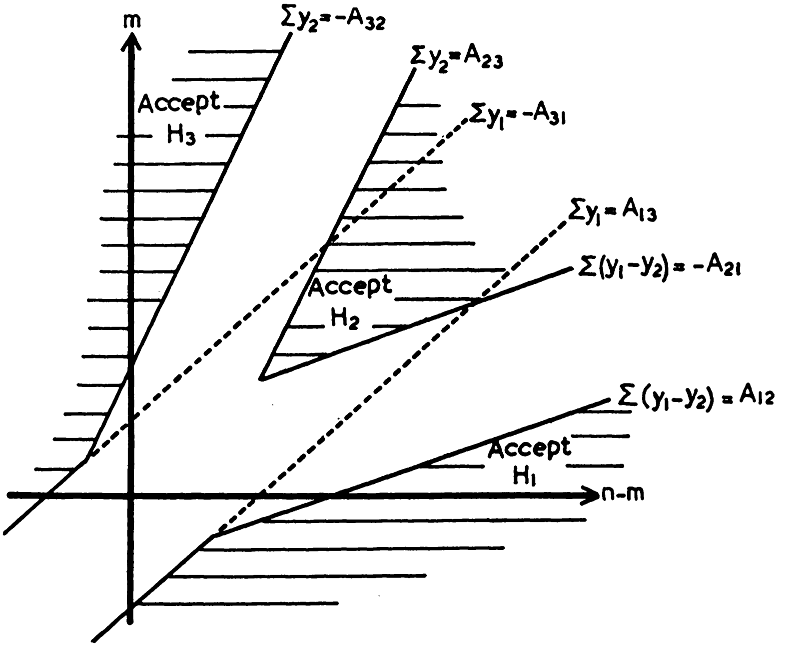


FIG. 3.—The inspection diagram in a sequential test for three binomal probabilities.

The only part of the inspection diagram which is used is, of course, the positive quadrant, and in Fig. 3 the boundaries of the acceptance regions for H_1 and H_3 are shown as being linear within this quadrant. This is not necessarily so.

The boundary $\Sigma y_2 = - A_{32}$ meets the m-axis where $m_2 = - A_{32}/\log (p_2/p_3)$, and the line

* *Cf.* Barnard (1946) or Stockman and Armitage (1946) (who use the term "lattice diagram"). Wald uses a similar diagram with n as abscissa and m as ordinate.

$\Sigma y_1 = -A_{31}$ meets the m-axis where $m_1 = -A_{31}/\log(p_1/p_3)$. In order that the two lines meet outside the positive quadrant we require

$$m_2 > m_1,$$

that is,

$$\frac{A_{32}}{A_{31}} > \frac{\log(p_3/p_2)}{\log(p_3/p_1)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The right-hand member of (6) is less than unity, and the inequality is satisfied if, for instance, $A_{31} = A_{32}$. The boundary of the acceptance region for H_3 is then linear. Similarly, the boundary of the acceptance region for H_1 is linear if

$$\frac{A_{13}}{A_{12}} < \frac{\log(1-p_1)/(1-p_3)}{\log(1-p_1)/(1-p_2)} \quad ,$$

which is certainly true if $A_{13} = A_{12}$, since $p_2 < p_3$.

It is interesting to note that the boundaries of Fig. 3 are almost the same as they would be if one were running two tests of the type considered by Wald in parallel—one for testing H_1 against H_2 and the other for testing H_2 against H_3. Such a combination of tests was suggested by Armitage (1947) as a sequential analogue of the two-sided t-test, the problem having been first reduced to a two-sided test of a binomial probability. The near equivalence of the two procedures is clear from the similarity between Fig. 2 of the 1947 paper and Fig. 3 of the present paper. (A slight difference is that the type of path exemplified by the dotted line on the 1947 diagram might, with a small probability, lead to the acceptance of H_1 or H_3 by the present procedure, whereas (in the present notation) H_2 would immediately be accepted by the previous procedure. This is clearly a very small discrepancy.)

This equivalence enables us to improve on the limits of error given by (4). Suppose, for simplicity, that $A_{ij} = A$, $B_{ij} = B$. Then, from (4),

$$\left.\begin{array}{l} \pi_{11} > 1 - 2/B \\ \pi_{22} > 1 - 2/B \\ \pi_{33} > 1 - 2/B \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

Now, if we were running two separate tests, with boundaries as in Fig. 3, and if in each test α were the probability that any hypothesis would be rejected when true, we should have (by Wald's theory)

$$B \doteq (1-\alpha)/\alpha, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7\text{A})$$

and (7) become, approximately,

$$\left.\begin{array}{l} \pi_{11} > 1 - 2\alpha/(1-\alpha) \doteq 1 - 2\alpha \\ \pi_{22} > 1 - 2\alpha/(1-\alpha) \doteq 1 - 2\alpha \\ \pi_{33} > 1 - 2\alpha/(1-\alpha) \doteq 1 - 2\alpha \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad (8)$$

if α is small.

But, by direct considerations (the argument here follows that of the 1947 paper), we see that the probability, in the combined test, of accepting H_1 when true is effectively the same as that of accepting H_1 when true in the separate test for H_1 against H_2, namely $1-\alpha$. Hence

$$\pi_{11} \doteq 1 - \alpha \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

and similarly

$$\pi_{33} \doteq 1 - \alpha \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

The probability, in the combined test, of rejecting H_2 when true is very nearly equal to the probability of rejecting H_2 when true in the test for H_1 against H_2, together with the probability of rejecting H_2 when true in the test for H_2 against H_3. Hence

$$\pi_{22} \doteq 1 - 2\alpha \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

A comparison between (8) and (9)–(11) shows that two of the three inequalities were unnecessarily wide. The reason is that, in this particular problem, when H_1 is true we are very unlikely to accept H_3, and *vice versa*, so that π_{13} and π_{31} are very small.

We may note, from the nature of Fig. 3, that the curve of the mean sample size against p will have a minimum when p is near p_2, and two maxima, one when p is between p_1 and p_2, and the other when p is between p_2 and p_3.

The relation between the present procedure and that given by Sobel and Wald (1949) becomes clear if we consider the following three mutually exclusive and exhaustive composite hypotheses:

$$H_1^*:\ p < a_1 \qquad H_2^*:\ a_1 \leqslant p \leqslant a_2 \qquad H_3^*:\ p > a_2,$$

where a_1 and a_2 are constants satisfying $p_1 < a_1 < p_2 < a_2 < p_3$.

A sequential procedure for deciding between H_1^*, H_2^* and H_3^* may be formulated by using exactly the same acceptance criteria as are given below equation (5), except that H_1^*, H_2^* and H_3^* replace H_1, H_2 and H_3 respectively. It follows from (9)–(11) that this decision procedure satisfies the following requirements:

(*a*) If $p \leqslant p_1$, the probability is $\geqslant 1 - \alpha$ that H_1^* will be accepted.

(*b*) If $p_1 \leqslant p \leqslant p_2$, the probability is $\geqslant 1 - \alpha$ that either H_1^* or H_2^* will be accepted, i.e. that H_3^* will be rejected.

(*c*) If $p = p_2$, the probability is $1 - 2\alpha$ that H_2^* will be accepted.

(*d*) If $p_2 \leqslant p \leqslant p_3$, the probability is $\geqslant 1 - \alpha$ that either H_2^* or H_3^* will be accepted, i.e. that H_1^* will be rejected.

(*e*) If $p \geqslant p_3$, the probability is $\geqslant 1 - \alpha$ that H_3^* will be accepted.

If, in the method of Sobel and Wald, we consider the particular case in which (in their notation) $\theta_2 = \theta_3$, and if we formulate an analogous problem in which the mean of a normal distribution is replaced by a binomial probability, we arrive at the above procedure, p_1, p_2 and p_3 being equivalent to θ_1, θ_2 and θ_3 respectively.

5. *A Two-sided Test for Comparative Trials*

Another problem which may be reduced to a sequential binomial test is the comparative trial for the difference between the parameters of two binomial distributions. Suppose we have two binomial distributions with parameters θ_1 and θ_2. In Chapter 6 of his book Wald (1947) suggests that successive observations from the two populations should be paired. Denoting an observation by 0 or 1, we consider only the pairs (0, 1) or (1, 0). If θ_1 and θ_2 are the probabilities that an observation will be a 1, the probability that a member of the sequence of pairs (0, 1) and (1, 0) will in fact be (0, 1) is

$$p = \frac{\theta_2(1 - \theta_1)}{\theta_1(1 - \theta_2) + \theta_2(1 - \theta_1)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

If $\theta_1 = \theta_2$, $p = \frac{1}{2}$. If $\theta_1 > \theta_2$, $p < \frac{1}{2}$; and if $\theta_2 > \theta_1$, $p > \frac{1}{2}$.

The hypothesis H_2 that $\theta_1 = \theta_2$ is therefore equivalent to the hypothesis that $p = \frac{1}{2}$. As alternative hypotheses about the "difference" between θ_1 and θ_2 we could conveniently choose the hypothesis H_1 that $p = p_1 < \frac{1}{2}$, and the hypothesis H_3 that $p = p_3 > \frac{1}{2}$. A simple hypothesis about p is a composite hypothesis about the original observations, since, for a given p, θ_1 is a function of θ_2 given by (12).

A test for the hypotheses H_1, H_2 and H_3 may now be constructed as in §4, the probabilities of correctly accepting each hypothesis being related to the positions of the boundaries by (7A) and (9)–(11). This test may in some practical situations be more useful than a test based on only two hypotheses. In a clinical trial to compare the effectiveness of two different treatments we may wish to design a test with the guarantees—

(*a*) that if one treatment is an improvement on the other by more than a certain amount ($p < p_1$ or $p > p_3$), we shall have a specified high probability of detecting the difference, and

(*b*) that if the treatments are equally effective, we shall have a specified high probability of saying so.

It may also be an advantage that the test has a comparatively small mean sample size when $p = \frac{1}{2}$, when $p < p_1$, or when $p > p_3$.

I wish to thank Mr. A. M. Walker for some helpful advice in the presentation of this paper, and Mrs. M. G. Young for preparing the diagrams.

References

Armitage, P. (1947), "Some sequential tests of Student's hypothesis," *J.R. Statist. Soc. Suppl.*, **9**, 250.
Barnard, G. A. (1946), "Sequential tests in industrial statistics," *J.R. Statist. Soc. Suppl.*, **8**, 1.
—— (1947), "Review of 'Sequential Analysis' by A. Wald," *J. Amer. Statist. Ass.*, **42**, 658.
Pearson, K. (1931), *Tables for Statisticians and Biometricians*, Part II. Cambridge University Press.
Rao, C. R. (1947), "The problem of classification and distance between two populations," *Nature, Lond.*, **159**, 30.
—— (1948), "The utilization of multiple measurements in problems of biological classification," *J.R. Statist. Soc.*, B, **10**, 159.
Smith, C. A. B. (1947), "Some examples of discrimination," *Ann. Eugen., Lond.*, **13**, 272.
Sobel, M., & Wald, A. (1949), "A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution," *Ann. math. Statist.*, **20**, 502.
Stockman, C. M., & Armitage, P. (1946), "Some properties of closed sequential schemes," *J.R. Statist. Soc. Suppl.*, **8**, 104.
Wald, A. (1947), *Sequential Analysis*. New York: John Wiley.
Welch, B. L. (1939), "Note on discriminant functions," *Biometrika*, **31**, 218